UPOTREBA ALGORITMA NAIVE BAYES U PROCESU DONOŠENJA POSLOVNIH ODLUKA

Marko Marković[1,*],
Katarina Plećić[1],
Živana Krejić[2],
Biljana Tešić[1]

[1]Fakultet zdravstvenih, pravnih i poslovnih studija Valjevo, Univerzitet Singidunum, Srbija
[2]Fakultet za poslovne studije i pravo, Univerzitet Union, Beograd, Srbija

Rezime:
Postoji značajan broj sistema za pomoć u poslovnom odlučivanju koji kao osnovu koriste Bajesovu teoremu. U ovom radu, prikazuje se jedna od primena ove teoreme korišćenjem algoritma Naive Bayes koji svoju upotrebnu vrednost dokazuje u mnogobrojnim praktičnim primenama upravo zbog jednostavnosti u fazama učenja i klasifikacije, kao i zbog činjenice da se može obučavati na izuzetno malim skupovima podataka i postići prilično veliku brzinu rada u poređenju sa drugim algoritmima mašinskog učenja. U ovom radu, prikazuju se matematičke osnove i način rada algoritma Naive Bayes, uz osvrt na skupove podataka sa kojima se pogodno koristi. Takođe, predstavlja se i praktično softversko rešenje koje su autori razvili za rad sa ovim algoritmom i kao pomoć u donošenju poslovnih odluka.

Ključne reči:
Bajesova teorema, Naive Bayes, poslovno odlučivanje.

1. UVOD

Naive Bayes modeli se često, i sa dosta uspeha, koriste za klasterovanje [1] i klasifikaciju [2]. U mnogim situacijama pokazuju izuzetno dobre karakteristike, a njihova preciznost i vreme učenja se mogu meriti i sa drugim modelima, kao što su Bajesove mreže [3], čiju jednu vrstu zapravo i predstavljaju. Čini ga mešavina komponenti, gde se u okviru svake komponente promenljive smatraju međusobno uslovno nezavisnim. Uz dovoljan broj podataka, može predvideti rešenje sa velikim stepenom preciznosti. Pritom, potreban skup podataka zapravo može biti prilično mali, zbog čega se ovaj algoritam dobro pokazao u situacijama kada ne postoji puno ulaznih podataka.

Algoritam *Naive Bayes*, kao i mnogi drugi sistemi koji koriste rasuđivanje zasnovano na verovatnoći, imaju za temelj Bajesovu teoremu [4]. Ovu teoremu je u XVIII veku razvio engleski matematičar, *Thomas Bayes*. Sam algoritam *Naive Bayes* je prvi put predstavljen tek početkom 1960.-ih godina, a dobio je ime po matematičaru na čijoj teoremi je zasnovan.

Bitno je pomenuti i poreklo naziva algoritma – iako bi neko iz naziva mogao da zaključi da se radi o „naivnom“ algoritmu, zapravo se

Odgovorno lice:
Marko Marković

e-pošta:
mmarkovic@singidunum.ac.rs

DOI: 10.15308/Sinteza-2019-215-220

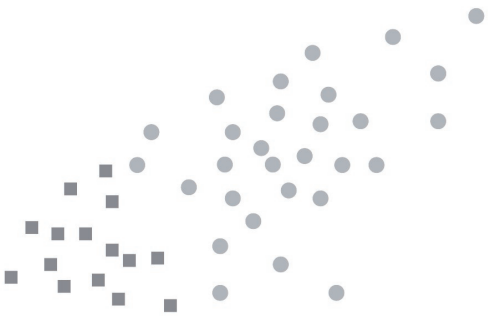
radi o tome da se podrazumeva da su atributi između sebe uslovno nezavisni [5]. I pored toga, u praksi, *Naive Bayes* modeli mogu raditi iznenađujuće dobro, u zavisnosti od prirode verovatnosnog modela, a do sada su svoju veliku upotrebnu vrednost dokazali u nekoliko veoma kompleksnih stvarnih situacija kao što je filtriranje spam poruka ili klasifikacija dokumenata [6]. U literaturi se povremeno ovi modeli nazivaju i "jednostavni Bajes" (*Simple Bayes*) i "nezavisni Bajes" (*Independence Bayes*) modeli.

2. NAČIN RADA ALGORITMA „NAIVE BAYES"

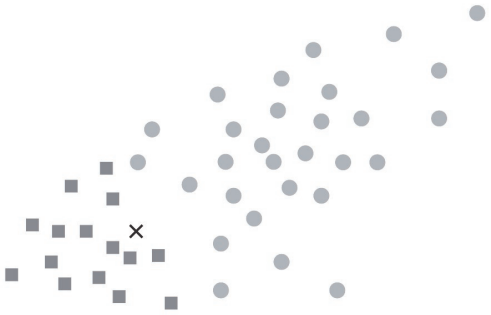
Algoritam Naive Bayes, kao i većina sistema koji koriste rasuđivanje zasnovano na verovatnoći, imaju za temelj Bajesovo pravilo (Bajesovu teoremu).

Kao demonstraciju osnovne ideje klasifikacije pomoću Bajesovog pravila možemo uzeti primer sa slike 1. Početne vrednosti podeljene su u dve kategorije i predstavljene su ili slikom kruga ili slikom kvadrata.

Potrebno je obezbediti da svaka nova vrednost koja se naknadno pojavi (slika 2) bude označena slikom kvadrata ili kruga. Trenutno, nova vrednost predstavljena je znakom X. Postavlja se pitanje na koji način je moguće odrediti u koju kategoriju će biti razvrstana. Jedna mogućnost je da bude označena kao kvadrat iz razloga što se nalazi bliže vrednostima tog tipa. Druga mogućnost je da bude označena kao krug jer je vrednosti koje su tako kategorizovane brojno dvostruko više. Verovatnoće događanja dva navedena scenarija su definisane na osnovu prethodnog iskustva (rasporeda i kategorizacije elemenata) i one se koriste za predviđanje novih ishoda korišćenjem Naive Bayes metoda.



Slika 1. Postojeće vrednosti u skupu podataka



Slika 2. Dodavanje nove vrednosti u postojeći skup podataka

Na osnovu prethodnog primera, može se definisati opšti slučaj Bajesovog pravila u sledećem obliku navedenom u izrazu (1).

$$P(b \mid a) = \frac{P(a \mid b)P(b)}{P(a)} . \quad (1)$$

Zahvaljujući Bajesovom pravilu, možemo izračunati P(b|a) pomoću P(a|b), P(b) i P(a). Iako, površno gledano, to ne izgleda naročito korisno, zapravo u praksi često postoje situacije kada se poznate ove tri vrednosti, a traži se četvrta.

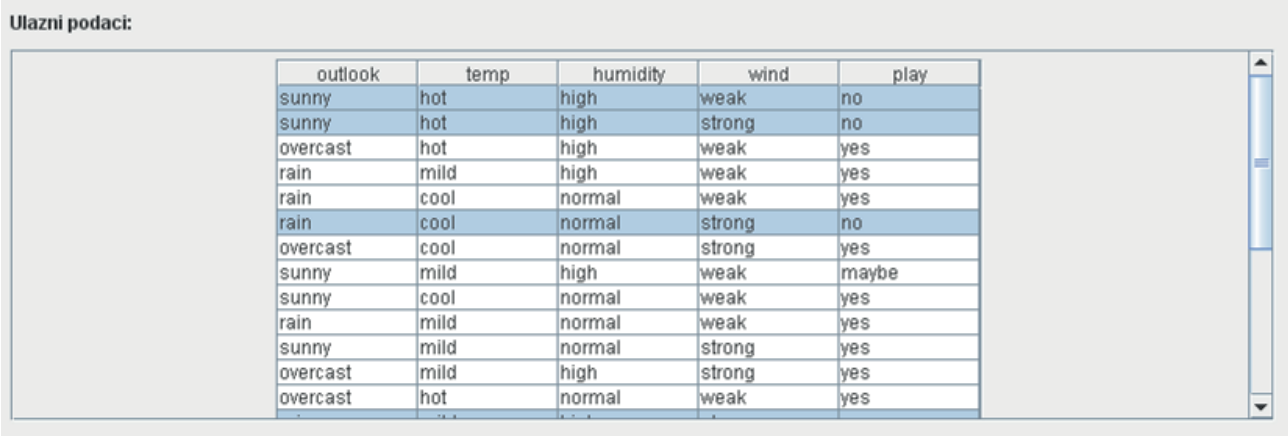
Bitno je pomenuti i poreklo naziva algoritma – iako bi se iz naziva moglo zaključiti da se radi o „naivnom" algoritmu, zapravo se radi o tome da se podrazumeva da su atributi između sebe uslovno nezavisni. Na taj način, svaki atribut doprinosi donošenju krajnje odluke u jednakoj meri i potpuno nezavisno od drugih atributa, što je efikasnije od drugih klasifikatora, gledano iz aspekta broja potrebnih operacija [7]. Zahvaljujući tome, Naive Bayes modeli u praksi mogu raditi iznenađujuće dobro.

S obzirom na navedenu pretpostavku o međusobnoj nezavisnosti atributa, verovatnoća posmatranja konjunkcije $a_1, a_2, \ldots, a_n$ je proizvod verovatnoća pojedinačnih atributa [8][9] kao što prikazuje (2).

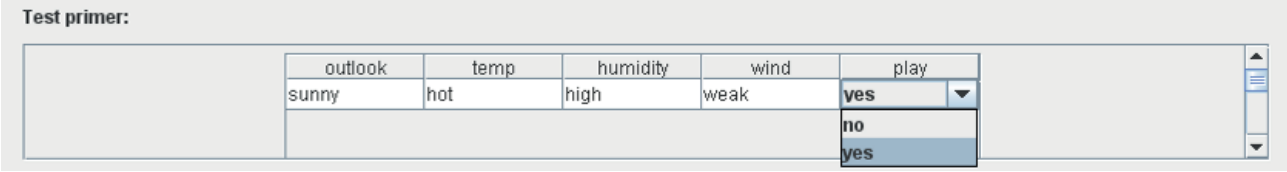
$$P(a_1, a_2, \ldots, a_n \mid v_j) = \prod_i P(a_i \mid v_j) \quad (2)$$

Algoritam *Naive Bayes* koristi ovu tvrdnju kao osnovu svog rada, a ona čini osnovu njegovog funkcionisanja i prikazana je u (3).

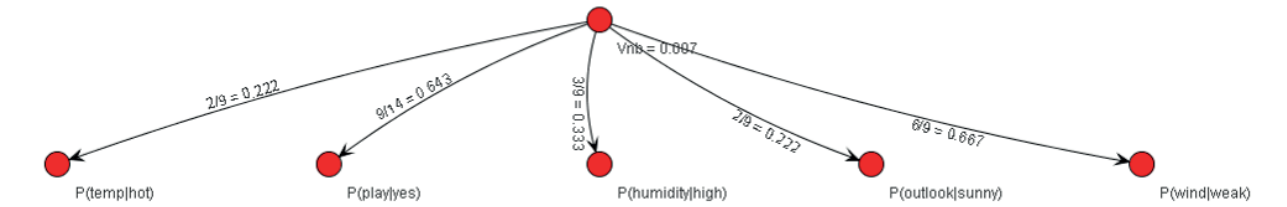
$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j) \quad (3)$$



Slika 3. Označavanje redova u tabeli sa ulaznim podacima



Slika 4. Izbor vrednosti test atributa



Slika 5. Prikaz stabla sa vrednostima uslovnih verovatnoća

Vrednost v_{NB} označava izlaznu vrednost algoritma. V predstavlja konačni skup ulaznih vrednosti, a njeni pojedinačni članovi obeleženi su sa v_j. Svaki primer iz skupa obučavanja predstavljen je kao skup vrednosti atributa (a_1, a_2, ..., a_n). Bitno je naglasiti da je broj različitih $P(a_i|v_j)$ uslova, koji moraju biti procenjeni iz skupa obučavanja, jednak broju različitih vrednosti atributa pomnoženi sa brojem različitih ciljnih vrednosti. [10] Takođe, ulazni skup podataka za algoritam koji se predstavlja u ovom radu mora sadržati diskretne vrednosti. U slučaju potrebe za kontinualnim vrednostima, potrebno je napraviti određene pretpostavke po pitanju distribucije vrednosti, a za tu svrhu se koriste druge varijacije algoritma koje nisu predmet ovog rada (na primer, alogirtam *Gausov Naivni Bajes – Gaussian Naive Bayes*).

Interesantna razlika algoritma Naive Bayes u odnosu na druge algoritme mašinskog učenja (na primer, stablo odlučivanja) je da ne postoji izričita pretraga prostora mogućih hipoteza. Umesto toga, hipoteza se formira bez pretraživanja, prebrojavanjem frekvencija raznih kombinacija podataka u okviru skupa obučavanja. [11]

Ovaj algoritam se pokazuje izuzetno efikasnim u širokoj lepezi primena, posebno zbog svoje jednostavnosti u fazi učenja i klasifikacije [12]. Izuzetno puno se koristi i u

oblasti klasifikacije teksta i slika, što je do sada potvrđeno u mnogim radovima i primerima iz prakse [13, 14]. Njegova važna karakteristika je da se dobro prilagođava i veoma velikim problemima: sa *n* Bulovih atributa, postoji samo *2n+1* parametara. Važno je naglasiti i da nema poteškoća sa nedostajućim podacima i podacima sa šumom. [10]

Čak se i situacije kada postoje nedostajuće vrednosti u ulaznim podacima ili test primerima mogu prilično jednostavno prevazići. Ukoliko u nekom redu sa ulaznim podacim nedostaje određena vrednost, ceo taj red se može ignorisati tokom ciklusa učenja algoritma. Moguće je čak nedostajuću vrednost ignorisati i u test primeru kako ne bi negativno uticala na krajnji rezultat.

Algoritam *Naive Bayes* poseduje i neke negativne strane na koje je potrebno obratiti pažnju prilikom izbora adekvatnog algoritma za potrebe donošenja poslovnih odluka. Upravo jedno od najvažnijih ograničenja predstavlja činjenica da se rad algoritma zasniva na pretpostavci da su atributi između sebe uslovno nezavisni - u realnom životu, gotovo je nemoguće pronaći skup atributa koji su međusobno potpuno nezavisni.

Takođe, ukoliko se u određenoj kategoriji pojavi vrednost koje nije bilo u skupi na kom se algoritam obučavao, biće dodeljena vrednost verovatnoće 0 – obzirom na pretpostavku uslovne nezavisnosti, kada se sve verovatnoće pomnože sa 0 i krajnja vrednost će imati tu vrednost – samim tim algoritam neće moći da napravi predviđanje. Ovaj problem se može dogoditi kada skup na kom se algoritam obučava ne predstavlja u potpunosti reprezentativan uzorak populacije. Postoje i određene tehnike koje se mogu koristiti da se ovakve situacije izbegnu – na primer, *Laplace smoothing*.

3. PRAKTIČNA PRIMENA ALGORITMA NAIVE BAYES

Za potrebe praktične primene algoritma Naive Bayes, autori su razvili softversko rešenje za pomoć u donošenju poslovnih odluka zasnovano na Java platformi. [15, 16] Softver omogućava da se unese proizvoljan ulazni skup podataka, da se iz posebne tabele izaberu test primeri koji se nalaze u odgovarajućim padajućim listama i da se algoritam potom obuči na ulaznom skupu i donese odluku na osnovu test primera. U svakom trenutku, korisnik može kontrolisati rad algoritma tako što će ga pauzirati ili će se kretati napred ili nazad kroz algoritam korak po korak i pratiti međufaze izvršavanja.

Za potrebe rada algoritma *Naive Bayes* može se koristiti bilo koji ulazni skup podataka, a za demostraciju u ovom radu koristi se *play tennis* primer (tabela 1). Ovaj skup podataka se često koristi kao demonstracija rada algoritama poslovnog odlučivanja. Navedeni skup podataka predstavlja vrednosti na osnovu kojih algoritam uči.

Tabela 1. Ulazni skup podataka primera play tennis

outlook	temp	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rain	mild	high	strong	no

Nakon što se vrednosti ovog primera učitaju u softver, prikazuju se u grafičkom interfejsu programa u tabeli prikazanoj na slici 3. U prikazu su trenutno označeni redovi čiji atributa *play* imaju vrednost *no*. Označene vrednosti se po potrebi mogu promeniti u skladu sa potrebama za lakšim pregledom korisnika koji upotrebljava program.

Nakon unosa ulaznih podataka, potrebno je izabrati vrednosti atributa za test primer koji će se koristiti (slika 4).

Algoritam radi tako što se za određenu kombinaciju atributa (u ovom slučaju vremenskih uslova) obračunava verovatnoća da li će se meč odigrati ili neće. Kao što je prikazano na slici 4, u ovom slučaju je potrebno odrediti verovatnoću da li će se teniski meč igrati, ukoliko je sunčano vreme, visoka temperatura i vlažnost vazduha, a vetar slab. Naravno, moguće je napraviti i bilo koju drugu kombinaciju test parametara.

Nakon pokretanja algoritma, vrši se obračun vrednosti i iscrtava se stablo sa prikazanim vrednostima uslovnih verovatnoća (slika 5).

U listovima stabla, nalaze se vrednosti uslovnih verovatnoća, kao i objašnjenje načina na koji su dobijene, tako da se može samostalno proveriti kako se došlo do određenih vrednosti. U korenu stabla nalazi se tražena verovatnoća da će meč biti odigran, uz zadate vremenske uslove iz test primera.

Slika 6 prikazuje tekstualni izlaz programa pomoću koga se može videti kako algoritam funkcioniše.

```
Korak 1
  P(play|yes) = 9 / 14 = 0.643
Korak 2
  P(outlook|sunny) = 2 / 9 = 0.222
Korak 3
  P(temp|hot) = 2 / 9 = 0.222
Korak 4
  P(humidity|high) = 3 / 9 = 0.333
Korak 5
  P(wind|weak) = 6 / 9 = 0.667
Korak 6
  Pronađeno rešenje
  Vnb = 0.007
```

Slika 6. Tekstualni izlaz algoritma Naive Bayes

Kao što se vidi iz korena stabla i tekstualnog izlaza algoritma, verovatnoća da će se igrati teniski meč je *0.007*. Da bi ova vrednost bila još korisnija, potrebno je izračunati verovatnoću da se pri zadatim vremenskim uslovima teniski meč neće odigrati. Ukoliko se ponovi prethodni postupak i za taj slučaj, dobiće se verovatnoća od *0.027*. Time zaključujemo da je veća verovatnoća da se teniski meč *neće* odigrati.

Od izuzetne važnosti je obratiti dodatnu pažnju na realizaciju procesa obračuna krajnje verovatnoće, a vezano za preciznost brojeva sa pokretnim zarezom. Naime, kada se uslovne verovatnoće (koje su uglavnom veoma mali brojevi) međusobno množe, mogu nastati još manje vrednosti – u tim situacijama postoji mogućnost gubitka preciznosti u računanju, o čemu se tokom razvoja softvera posebno mora voditi računa.

4. ZAKLJUČAK

U oblasti primene inteligentnih sistema u poslovnom odlučivanju postoji veliki broj algoritama koji se mogu koristiti kao pomoć pri donošenju odluka. U okviru ovog rada, cilj je bio da se predstavi Naive Bayes kao jedan od algoritama za koji su autori smatrali da sa jedne strane omogućava brz i efikasan proces učenja i klasifikacije, a da sa druge strane nudi značajnu prednost po pitanju kvaliteta odluka koje predviđa. Ovaj algoritam je i specifičan po tome što poseduje izuzetno veliku brzinu rada. To znači da se u nekim slučajevima verovatnoće mogu izračunavati čak i kod svake promene ulaznih podataka, što je nezamislivo za mnoge druge algoritme mašinskog učenja.

Kada su ulazni podaci takvi da se može održati pretpostavka što višeg stepena uslovne nezavisnosti atributa, mogu se ostvariti i bolji rezultati u odnosu na druge modele, a često i ulazni skupovi podataka na kojima algoritam uči mogu biti manji. Iako postoje i određeni nedostaci, uz adekvatnu pažnju i primenu korektivnih mera, mogući problemi se mogu izbeći. Upravo korišćenjem softvera koji su razvili autori rada, moguće je primeniti *Naive Bayes* algoritam, ali i kroz grafički izlaz programa lako uvideti sve uslovne verovatnoće koje čine krajnji rezultat. Zahvaljujući tome, nastao je alat koji omogućava lako korišćenje algoritma, čak i za one korisnike koji nisu detaljno upućeni u sam princip rada navedenog algoritma.

ZAHVALNOST

Ovaj rad je proistekao iz rezultata rada na projektu Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije, broj TR32054.

LITERATURA

[1] P. Cheeseman and J. Stutz, „Bayesian classification (AutoClass): Theory and results.“ In Advances in knowledge discovery and data mining, AAAI Press, Menlo Park, CA, 1996.

[2] P. Domingos and M. Pazzani, „On the optimality of the simple Bayesian classifier under zero-one loss.“ Machine Learning, 29, 1997.

[3] D. Lowd and P. Domingos, „Naive Bayes Models for Probability Estimation“, Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.

[4] E. M. Gredler, „Games and Simulations and their Relationships to Learning“, Handbook of Research on Educational Communications, Lawrence Erlbaum Associates Publishers, Mahwah, NJ, 2004.

[5] J. E. Dittrich, „Realism in Business Games: A Three Game Comparison“, University of Kentucky, Lexington, Kentucky, 1977.

[6] M. Wawer, M. Milosz, P. Muryjas and M. Rzemieniak, „Business Simulation Games in Forming of Students' Enterpreneurship“, International Journal of Euro-Mediterranean Studies, Volume 3, EMUNI University i University of Nova Gorica, Portorož, 2010.

[7] S. L. Ting, W. H. Ip and A. Tsang, „Is Naive Bayes a Good Classifier for Document Classification?“, International Journal of Software Engineering and Its Applications, Vol. 5, No. 3, July, 2011.

[8] P. A. Flach and N. Lachiche, „Naive Bayesian Classification of Structured Data“, Machine Learning, Volume 57, Issue 3, December 2004.

[9] J. Anderson, „Diskretna matematika sa kombinatorikom“, Pearson Education, Beograd, 2005.

[10] S. Russell and P. Norvig, „Veštačka inteligencija – savremeni pristup“, CET, Beograd, 2011.

[11] T. Mitchell, „Machine Learning“, McGraw Hill, Boston, 1997.

[12] S. Chakrabarti, S. Roy and M. V. Soundalgekar, „Fast and accurate text classification via multiple linear discriminant projection“, The VLDB Journal The International Journal on Very Large Data Bases, 2003.

[13] T. Joachims, „Text categorization with support vector machines: Learning with many relevant features“, In Proceedings: Machine Learning: ECML - 98, 10th European Conference on Machine Learning, 1998.

[14] S. McCann and D. G. Lowe, „Local Naive Bayes Nearest Neighbor for Image Classification“, CoRR, 2011.

[15] M. Marković, O. Nikolić and B. Nikolić, „Realizacija softverskog sistema za vizuelnu simulaciju algoritama mašinskog učenja“, Konferencija Sinteza, Univerzitet Singidunum, 2014.

[16] M. Marković, I. Kostić Kovačević, O. Nikolić and B. Nikolić, "INSOS—Educational System for Teaching Intelligent Systems", Computer Applications in Engineering Education, Wiley Blackwell, 2014.